# An Integrative Bioinformatics Framework for Linking Peripheral Biomarker Genes with Coronary Artery Disease

**Ranjan K. Pradhan**[1,2]
[1]Department of Biotechnology, Department of Electrical Engineering, College of Engineering and Technology, Bhubaneswar, Odisha, India
Email of Corresponding Author: rkpradhan@cet.edu.in

**Abstract**:
Coronary artery disease is the leading cause of mortality and morbidity worldwide and shows significant biomarker diversity. Despite considerable progress in microarray and proteomics technologies, and therapeutic strategies, the genetic basis of this disease remain poorly understood. High-throughput experiments such as genome-wide association studies and microarray gene expression profiling offer valuable information to uncover potential biomarkers associated with biochemical pathways and underlying mechanisms. Meta-analysis and gene enrichment analysis are known to provide powerful tool to discover novel biomarkers and their associated pathways using available genomic data on differentially expressed genes, measured under diverse conditions. Although meta-analysis alone has shown to predict biomarker genes of cardiovascular diseases, its efficacy in linking differentially expressed genes to coronary artery disease pathways along with other functional enrichment tools is unclear. Specifically, identifying differentially expressed genes using meta-analysis is often complicated by several factors such as the differences in microarray technique, study focus, data quality, gene nomenclature, species, intervention, and the statistical models used for analysis. Therefore, in this work, we developed a combined meta-analysis of multiple microarray data with quality control analysis, gene ontology and pathways functional enrichment analysis to identify potential differentially expressed genes of coronary artery disease and their cellular pathways. Gene expression data from four different studies were analyzed individually and using meta-analysis of combined data, which results in five differentially expressed genes that are thought to involve in coronary artery disease. This framework demonstrated the significance of microarray data heterogeneity and selection of statistical measures in identifying potential biomarkers associated with coronary artery disease, and can help elucidating the molecular mechanisms underlying this disease.

**Keywords**—*Microarray Data; Differentially Expressed Genes; Gene Enrichment; Bioinformatics Analysis; Meta-analysis; Coronary Artery Disease*
**DOI Number: 10.48047/nq.2019.17.08.2023** **NeuroQuantology 2019; 17(08):66-73**



## Introduction

Coronary artery disease (CAD) refers to the building-up of atherosclerotic plaque in the blood vessels that supply oxygen and nutrients to the heart[1]. Although multiple genetic and environmental factors and their interactions are known to affect the progression of CAD, current understandings of the molecular mechanisms of CAD and its diagnostic strategies remain incomplete [2]. Peripheral blood gene expression profiles have been extensively used to study pathological states in a variety of diseases including CAD [3-12]. Thus, identifying differentially expressed genes from a variety of microarray studies may help to elucidate the underlying molecular mechanisms, and discover novel biomarkers for improved therapeutic plans for CAD.

Microarray gene expression profiling is a powerful tool for studying genetic origin of various cardiovascular diseases, but the ability to compare the expression data and derive conclusions often found problematic [13, 14]. This is specifically due to the inherent biological, experimental, and methodological variations associated with different studies. Many of these can be overcome by using standard reporting methods, together with careful application of large-scale meta-analysis techniques, which is currently lacking for most disease genes[15]. Meta-analysis is a statistical tool that combines data from numerous studies, minimizes bias and increases statistical power by increasing sample size compared to individual studies.

The present study performed the meta-analysis of microarray datasets of four different studies that reported the changes in gene expression values from normal to disease conditions, and identified key gene targets by combined p-values. A protein-protein interactions network was reconstructed and functional gene enrichment analysis was done to identify key proteins involved in RNA and energy metabolism. The reactome pathway analysis revealed various enriched pathway terms. However, the present systematic analysis is able to reveal the limitation and strength of meta-analysis for microarray data mining. This study suggests the necessary methodological precautions that one must consider in applying meta-analysis, and try to improve the reliability and generalizability of this integrative method for identification of up- and down-regulated genes of multiple microarray studies related to coronary artery disease.

The paper is organized as follows. The meta-analysis workflow and its systematic application to different CAD data are described in section II. Section III presents the original data preparation, preprocessing, individual data analysis using t-test statistics and combined data analysis using meta-analysis tool, protein- protein interaction network analysis, reactome database analysis and gene enrichment analysiss results. Section IV discusses the different aspects of this meta-analysis tool in identifying candidate marker gene for CAD target. Finally, in Section V, it presents the general conclusions drawn from this analysis, limitation of the study and future scopes.

**Materials and Methods**

In this study, microarray gene expression profile data were analyzed using different quality control and statistical bioinformatics tools and databases. The proposed bioinformatics and data mining work flow, preprocessing techniques, quality control measures, and meta-analysis results were validated by examining the resulting performance indices for each single study. A systematic analysis and interpretation of biomarker genes and their associated pathways were presented as follows.

*Microarray Gene Expression Data of CAD Patients*

Microarray gene expression data from coronary artery disease genomic studies [3, 16-19] were used in present analysis using R-Studio and bioinformatics open source enrichment tools. Four gene expression datasets, including GSE20680, GSE20681, GSE42148 and GSE48060 were collected from the NCBI GEO database. These datasets were first analyzed using various preprocessing techniques and quality control methods. After verifying the intermediate samples, a total of 238 CAD cases and 189 matched or partially matched control cases were chosen for quality assessments and DEG analysis. A brief description of data size these of these five microarray datasets was summarized in Table 1. After pre-processing the number of data per study and number of sample were shown in the bar plot of Figure 2. Note that for each study, the datasets were downloaded as the series matrix files. Then series matrix files were extracted to text format for analysis.



Table 1: List of GEO data on Coronary Artery Disease genes.

| GEO Accession No | Platform | Control | Case | Microarray |
|---|---|---|---|---|
| GSE20680 | GPL4133 | 99 | 99 | Affymetrix |
| GSE20681 | GPL4133 | 52 | 87 | Affymetrix |
| GSE48060 | GPL570 | 21 | 31 | Affymetrix |
| GSE42148 | GPL13607 | 11 | 13 | Affymetrix |

In GSE20680 dataset, there are 52 controls and 87 CAD cases after omitting the intermediate values from the original raw data. In GSE20681 dataset there are 99 controls and 99 CAD cases. In GSE29532, there are total 55 samples were reported at different time points, but samples at the time of admission of the patients were taken for consideration and accordingly this data has only 6 controls and 8 CAD cases. Similarly, in GSE42148, there are 11 controls and 13 CAD cases were reported. In GSE48060, there are 21 controls and 31 CAD cases. Note that, in all the data sets, controls are the samples measured in healthy persons whereas CAD cases are the samples from unhealthy or diseased persons. Fig. 1 shows the proposed workflow in identifying the genes that were differentially expressed in individual data and combined data. Other detailed information of the four microarray datasets was summarized in Table 1 and shown in Fig. 2.



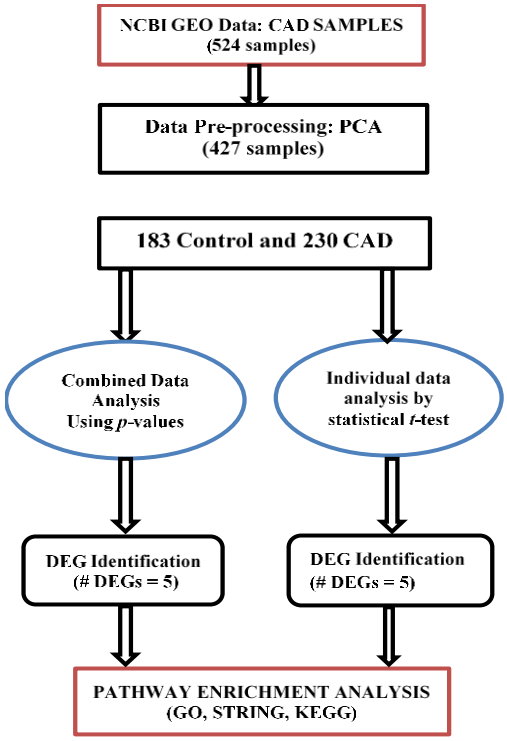


**Fig. 1:** Schematics of the bioinformatics workflow for identifying differentially expressed genes associated with coronary artery disease and their associated pathways.

*Data Pre-processing and Quality Assessment*

After individual data pre-processing, four different quality control measures were computed using MetaQC [20] to define the quality of these datasets. These measures includes, the internal quality control index (IQC), external quality control index (EQC), accuracy quality control index for genes (AQCg) and consistency of differential expression quality control (CQCg) index. IQC represented the internal homogeneity of co-expression. Here the EQC index was calculated based on an external pathway database MSigDB. AQC and CQC were basically meant for assuring of getting similar differentially expressed genes detected in an individual study compared to those detected by systematic analysis of other studies. For dimension reduction, Principal Component Analysis (PCA) was conducted and MetaQC was used to visualize the quality of individual data set. Figure 2 shows 96.7% of genes are along principal components 1and 2.

The four quality control measures were projected into a 2D space where the coordinates of each quality measure was determined by its correlation to first two principal components.

*Identification of DEGs*

For identification of differentially expressed genes, MetaDE package (R Studio) was used for systematic analysis of each dataset [20] and Fisher, adaptively weighted Fisher (AW), minimum p-value (minP), maximum p-value (maxP). Moderated t-test was used to decide the differentially expressed genes for a single dataset. The heat map of the differentially expressed genes under 0.15 FDR threshold across

studies were created. To evaluate the performance of these methods, we compared the numbers of detected DEG from different methods under different p-value thresholds using detection competency curves.

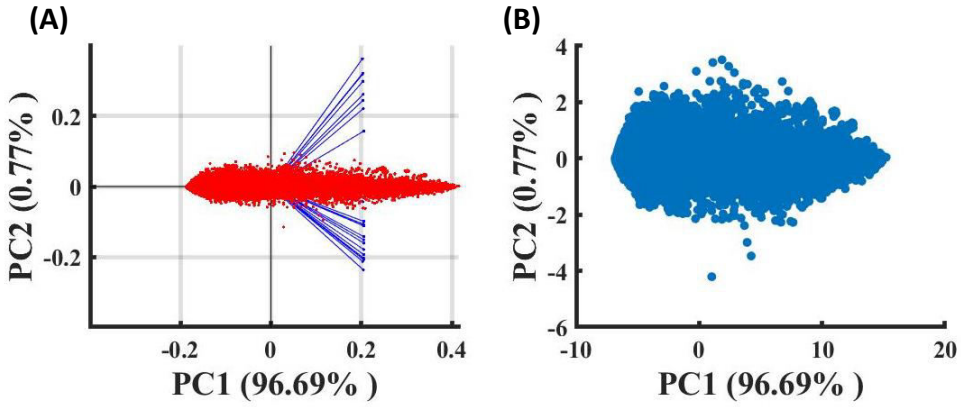


**Fig. 2:** Shown here is the (A) PCA biplot and (B) score plot of a sample microarray data (GSE20680) that suggests 97% of genes are along the PC1 and PC2.



*Gene-Enrichment Analysis*

The functional gene enrichment and pathway enrichment were carried out using Reactome database and FDR adjustment was applied to identify significantly enriched pathways [20]. The pathways shared by at least two datasets were plotted as heat map. Protein-protein interactions of the identified differentially expressed genes were also analyzed using STRING software tool that can show known and predicted protein-protein interactions.

**Results**

After the datasets were collected from NCBI Database and they were pre-processed in R-studio, which includes three key steps: $Log_2$ transformation, gene merging and gene filtering. This ensured that 20% unexpressed and 20% non-informative genes are removed, that could have increased false positive. Principal component analysis was used to visualize the data in low dimension and it was found that almost 91% of the data lies along the first two PCs. Based on these six quality control criteria and limited data available, we chose to include all four data sets into our framework.

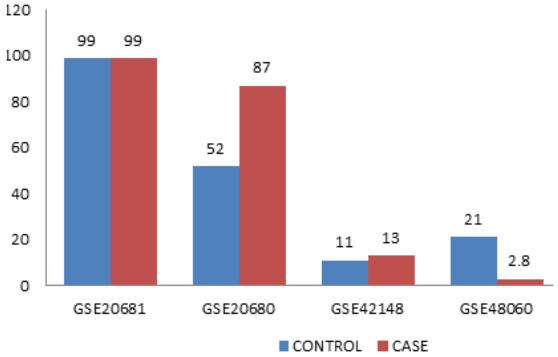


**Fig. 3**: The bar graph representing the number of controls and cases in the four datasets of coronary artery disease after pre-processing.

To find DEGs for single dataset, Reactome analysis tool was used. First, individual GEO data were analyzed and we were able to find the top up-regulated and down-regulated genes, as shown in Figure 4 for GSE20681 and GSE42148. For each individual data, effect of varying p-values and FDR on number of differentially expressed genes were analyzed. As evident from Table 2 that by setting different p-values, the number of DEGs found to vary, depending on FDR values.

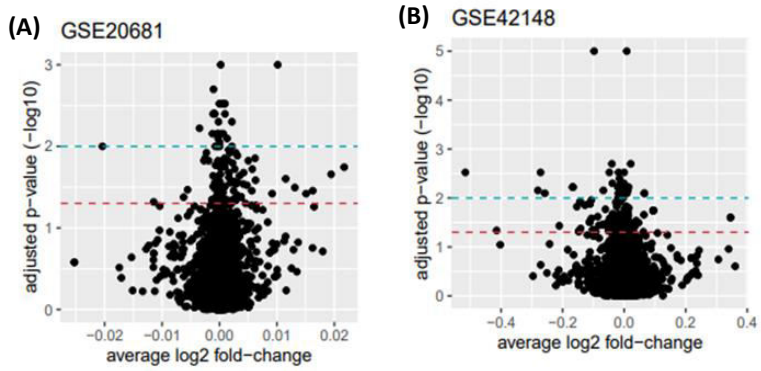


**Fig. 4**: Shown here are the adjusted *p*-values vs. average log2 fold changes of gene expression in GSE20681 and GSE 42148 showing the up- and down-regulated genes in CAD.

For identifying differentially expressed genes in combined data, R-based MetaDE package was used [20]. Five statistical measures based on computed p-value were used such as: Fisher, adaptively weighted Fisher (AW), minimum p-value (minP), maximum p-value (maxP) and rth-ordered p-value (roP). These five statistical methods were used by combining all p-values of Table 2. Total five differentially expressed genes were detected by maxP and roP evaluation criteria, respectively, using false discovery rate (FDR) cut-off less than 0.05. By setting FDR cut-off to 0.1, total 7 genes were differentially expressed. When FDR was set less than 0.5, we got 5 differentially expressed genes: C10RF213, CA7, TAF1C, TTC5, and PLA2G2A.

70

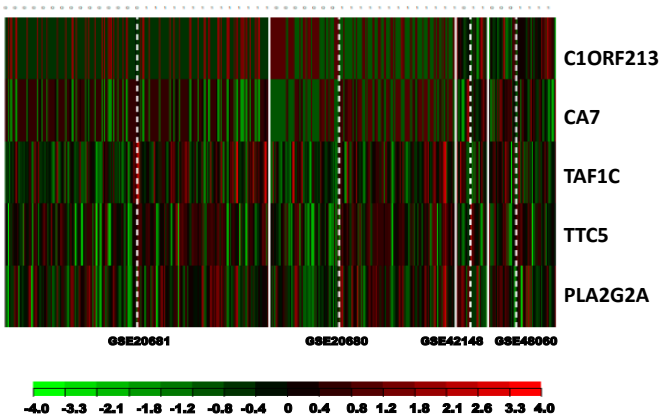


**Fig 5.** Heat-map identifying the differentially expressed genes (C10RF213, CA7, TAF1C, TTC5, PLA2G2A) in control and CAD cases, using Fisher meta-analysis method when FDR set to 0.05.

Table 3: Differentially expressed genes from four different CAD datasets using moderated t-test and meta-analysis combined p-value.

| *p*-value | GEO Data | | | | Method of Analysis | | | |
|---|---|---|---|---|---|---|---|---|
| | GSE 20681 | GSE 20680 | GSE 42148 | GSE 48060 | Fisher | MaxP | roP | AW |
| *p*=0.01 | 71 | 103 | 260 | 0 | 75 | 0 | 0 | 130 |
| *p*=0.05 | 437 | 447 | 974 | 0 | 294 | 0 | 0 | 492 |
| *p*=0.1 | 884 | 884 | 1669 | 0 | 554 | 0 | 0 | 911 |
| *p*=0.5 | 4953 | 4704 | 6145 | 0 | 3373 | 0 | 0 | 438 |

Gene enrichment analysis or functional analysis are critical to understand how a set of DEGs work. These genes may have been determined by an analysis of differential expression analysis, GWAS analysis, or a proteomics analysis. No matter where the gene list comes from, functional analysis can show if certain pathways or processes are enriched among a list of genes. Therefore, in this study we used open source Reactome analysis tools to study gene-gene interaction and their association with various pathways.

We analyzed the list of the significant genes that we obtained from the heat map. We found several significant pathways. Two data sets GSE20681and GSE42148 compared using Reactome analysis tool, which reveals 2479 pathways and 13561 fold changes in genes or proteins in the first data set (GSE20681), and 2482 pathways and 14381 fold changes for genes or proteins in the second dataset (GSE42148). Figure 5 shows the volcano plots, which summarize the pathway results for each dataset. Every point represents one pathway. The x-axis represents the average fold-change of all genes/proteins within that pathway. The y-axis represents the p-value where "higher" values are more significant (-log10 transformation). The red line represents p = 0.05, and the blue line p = 0.01. Note that the pathway correlation between GSE20681 and GSE42148 was obtained in form of a graph and only pathways that were observed in both datasets were shown. The volcano plot for the two datasets: GSE20681 (top plot) and GSE42148 (bottom plot), showing the statistically significant sets of genes with large fold changes.

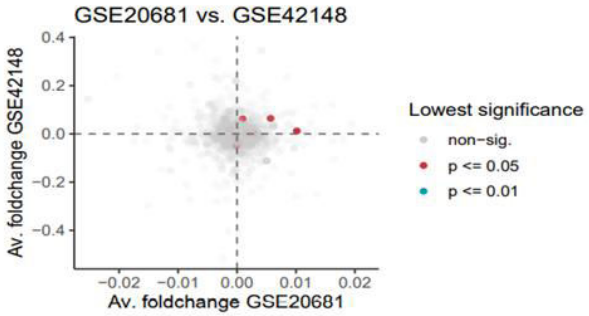


**Fig. 6**: The volcano plot showing the pathway correlation between the two datasets GSE20681 and GSE42148.



The correlation between various pathways of these datasets were computed, and shown in Figure 5. Note that, the x-axis is showing the average fold-change of one dataset and the y-axis is showing the average fold-change of the other dataset. Only pathways that were observed in both datasets are shown in this plot. Points are colored based on the lowest observed significance. The five DEGs identified from meta-analysis were taken for further gene enrichment analysis (in ShinyGO) to find out top 20 enriched GO terms and KEGG pathways as shown in Figure 7A. This results a pathway network showing the extent of overlapping and possible protein-protein interactions involved in CAD, as shown in Figure 7B.

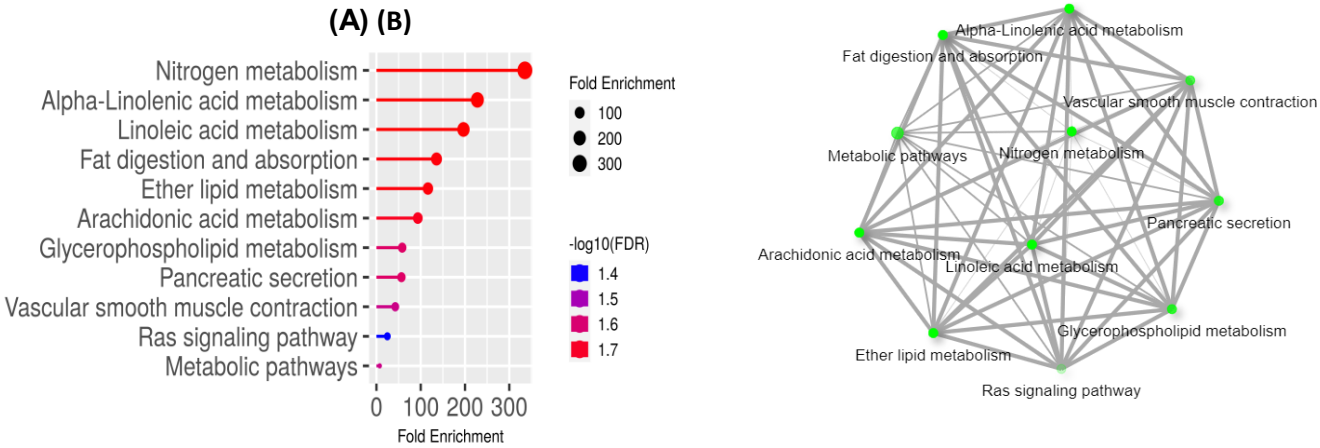

**Fig. 7:** Shown here is the (A) top 20 enriched pathways of KEGG database associated with five DEGs, with FDR cutoff as 0.05, and (B) a possible etwork of pathway interactions based on enrichment analysis of five DEGs identified in this study.

**Conclusion**

This study applied an integrated bioinformatics analysis by combining meta-analysis with DEG enrichment analysis, applied to four independent sets of microarray gene expression data on coronary artery disease. The effects of various statistical parameters, data quality and data preprocessing methods on the number of differentially expressed genes (DEGs) were explored to identify an optimal set of DEGs associated with CAD. Both individual data analysis and combined data analysis were conducted to find out the up-regulated and down-regulated genes and their associated biochemical pathways.

Meta-analysis of four data sets give five DEGs (C10RF213, CA7, TAF1C, TTC5, and PLA2G2A) for a FDR value fixed at 0.5. It was observed that the actual number of DEG largely depends on the overall p-values and FDR, in a given dataset. The present analysis revealed five different DEGs that were analyzed using reactome database tool, were found to be associated with 25 cellular pathways of immune system. This is consistent with previously studied results where it was extensively demonstrated that innate and adaptive immune responses have definite roles in the development and progression of cardiovascular disease. Also atherosclerosis, the main cause of coronary artery disease has been widely accepted as a chronic inflammatory disease. This is consistent with our results. However, the number of pathways was found to be higher for the case of individual data analysis (e.g., for GSE28061 and GSE42148, 2479 pathways and 2482 pathways were found). Thus, meta-analysis of micro array gene expression data is an effective way to determine differentially expressed genes within multiple studies, regardless of data size and methods used for measurement.

One important aspect of this study is that by adjusting few parameters it is possible to improve the statistical significance of predicted DEGs. However, the functional significance of DEGs needs further validation with additional data on same disease. As seen here this was overcome by iterative analysis of multiple datasets through combined meta-analysis and pathway enrichment of DEGs. Furthermore, the genetic processes underlying CAD have been widely demonstrated using a variety of -omics techniques, including genome-wide association studies, candidate gene studies and microarray experiments of differential gene expressions in CAD samples compared to controls. Although these studies produce highly valuable and informative data showing a snapshot of most of the genetic events occurring in a disease at a given point of time, their interpretation remain extremely challenging. In this view, the present study try to clarify the strength and weakness of four CAD datasets that are often used for biomarker identification and interpreting disease mechanisms.

Another striking feature of this study is to address some of the key concerns about the possibility of identifying a unique set of DEGs through meta-analysis. Specifically, it was illustrated how the results of the meta-analysis-based biomarker identification from expression data may significantly vary, depending on the sample size, number datasets available, particular clinical symptoms of patients, matched control conditions between multiple studies, and statistical methods used for analysis. However, adding gene enrichment analysis it is possible to eliminate potential source of variability in the prediction results on the number of biomarker genes that are differentially expressed. Our approach therefore uses simple criteria to select the common DEGs, followed by gene and pathway enrichment analysis. It was observed that although it is possible to derive CAD pathway information from publicly available expression data, integrated meta-analysis approach used in this study requires additional data for validating the involvements of the estimated set of DEGs in CAD, which is one of the future scopes of this study. However, the present integrative analysis highlights the importance of microarray gene expression data sharing in the public domain with minimal information bias.



**References**

[1] N. Poulter, “Coronary heart disease is a multifactorial disease,” *American Journal of Hypertension*, vol. 12, no. S6, pp. 92S-95S, Oct. 1999, doi: 10.1016/S0895-7061(99)00163-6.

[2] A. A. Vinkhuyzen, N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher, "Estimation and partition of heritability in human populations

using whole-genome analysis methods," Annu Rev Genet, vol. 47, pp. 75-95, 2013, doi: 10.1146/annurev-genet-111212-133258.

[3] R. Do et al., "Common variants associated with plasma triglycerides and risk for coronary artery disease," Nat Genet, vol. 45, no. 11, pp. 1345-52, Nov 2013, doi: 10.1038/ng.2795.

[4] D. Glass et al., "Gene expression changes with age in skin, adipose tissue, blood and brain," Genome Biol, vol. 14, no. 7, p. R75, Jul 26 2013, doi: 10.1186/gb-2013-14-7-r75.

[5] R. McPherson, "A gene-centric approach to elucidating cardiovascular risk," Circ Cardiovasc Genet, vol. 2, no. 1, pp. 3-6, Feb 2009, doi: 10.1161/CIRCGENETICS.109.848986.

[6] K. D. Coon, T. L. Dunckley, and D. A. Stephan, "A generic research paradigm for identification and validation of early molecular diagnostics and new therapeutics in common disorders," Mol Diagn Ther, vol. 11, no. 1, pp. 1-14, 2007, doi: 10.1007/BF03256218.

[7] M. A. Perera et al., "Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study," Lancet, vol. 382, no. 9894, pp. 790-6, Aug 31 2013, doi: 10.1016/S0140-6736(13)60681-9.

[8] S. Ripke et al., "Genome-wide association analysis identifies 13 new risk loci for schizophrenia," Nat Genet, vol. 45, no. 10, pp. 1150-9, Oct 2013, doi: 10.1038/ng.2742.

[9] I. E. Jansen et al., "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk," Nat Genet, vol. 51, no. 3, pp. 404-413, Mar 2019, doi: 10.1038/s41588-018-0311-9.

[10] K. D. Coon et al., "A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease," J Clin Psychiatry, vol. 68, no. 4, pp. 613-8, Apr 2007, doi: 10.4088/jcp.v68n0419.

[11] A. Huertas-Vazquez et al., "Novel loci associated with increased risk of sudden cardiac death in the context of coronary artery disease," PLoS One, vol. 8, no. 4, p. e59905, 2013, doi: 10.1371/journal.pone.0059905.

[12] G. T. Jones et al., "A sequence variant associated with sortilin-1 (SORT1) on 1p13.3 is independently associated with abdominal aortic aneurysm," Hum Mol Genet, vol. 22, no. 14, pp. 2941-7, Jul 15 2013, doi: 10.1093/hmg/ddt141.

[13] F. Cordero, M. Botta, and R. A. Calogero, "Microarray data analysis and mining approaches," Brief Funct Genomic Proteomic, vol. 6, no. 4, pp. 265-81, Dec 2007, doi: 10.1093/bfgp/elm034.

[14] N. E. Olson, "The microarray data analysis process: from raw data to biological significance," NeuroRx, vol. 3, no. 3, pp. 373-83, Jul 2006, doi: 10.1016/j.nurx.2006.05.005.

[15] T. C. Chalmers, "Meta-analysis in clinical medicine," Trans Am Clin Climatol Assoc, vol. 99, pp. 144-50, 1988. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/3332517.

[16] M. R. Elashoff et al., "Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients," BMC Med Genomics, vol. 4, p. 26, Mar 28 2011, doi: 10.1186/1755-8794-4-26.

[17] S. Dandona et al., "The transcription factor GATA-2 does not associate with angiographic coronary artery disease in the Ottawa Heart Genomics and Cleveland Clinic GeneBank Studies," Hum Genet, vol. 127, no. 1, pp. 101-5, Jan 2010, doi: 10.1007/s00439-009-0761-3.

[18] P. Beineke et al., "A whole blood gene expression-based signature for smoking status," BMC Med Genomics, vol. 5, p. 58, Dec 3 2012, doi: 10.1186/1755-8794-5-58.

[19] R. Suresh et al., "Transcriptome from circulating cells suggests dysregulated pathways associated with long-term recurrent events following first-time myocardial infarction," J Mol Cell Cardiol, vol. 74, pp. 13-21, Sep 2014, doi: 10.1016/j.yjmcc.2014.04.017.

[20] D. D. Kang, E. Sibille, N. Kaminski, and G. C. Tseng, "MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis," Nucleic Acids Res, vol. 40, no. 2, p. e15, Jan 2012, doi: 10.1093/nar/gkr1071.